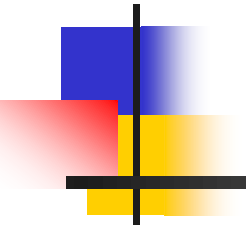
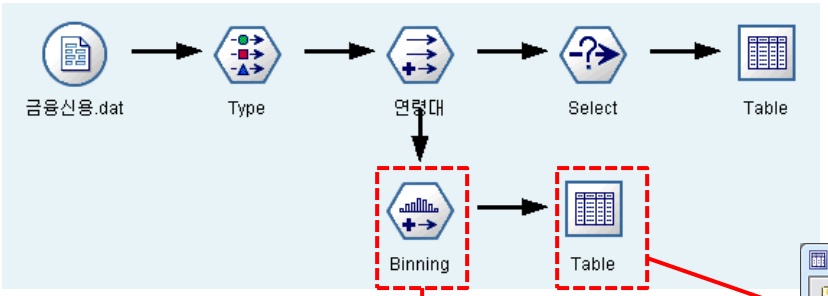


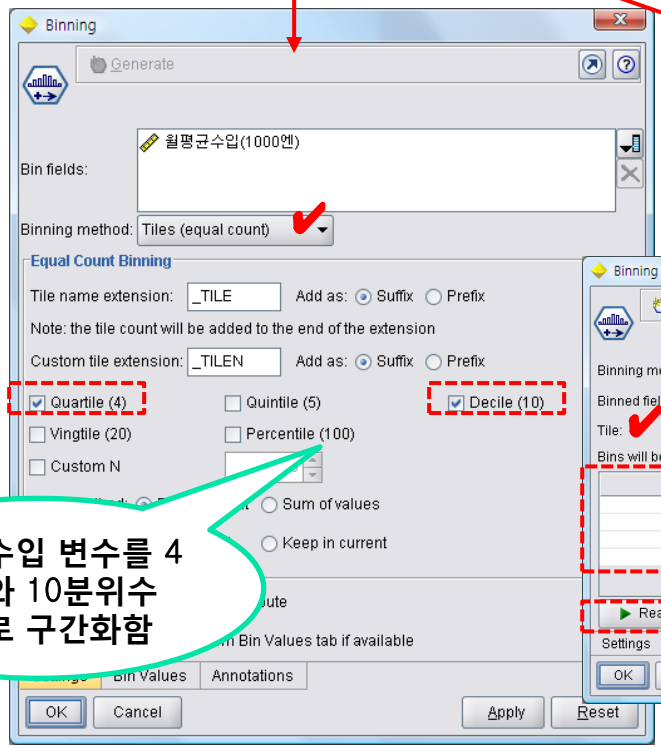
Data Munging with SPSS Modeler

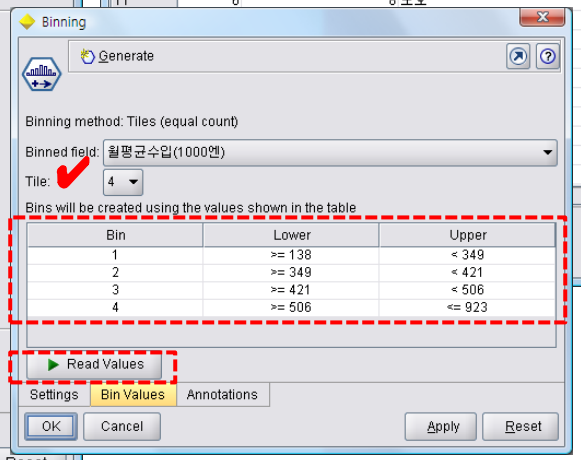
데이터 구간화

❖ **구간화의 필요성**
- 연속형 데이터의 범주화
- 범주화를 통해서 정확한 의미 부여
- 각종 분위수의 산정을 통한 정보의 이해 증진

❖ **구간화(Binning) 노드**
연속형 변수를 등간격, 등비율, 평균/표준편차 등을 고려한 형태의 구간화 규칙을 적용하여 범주형 변수로 변환



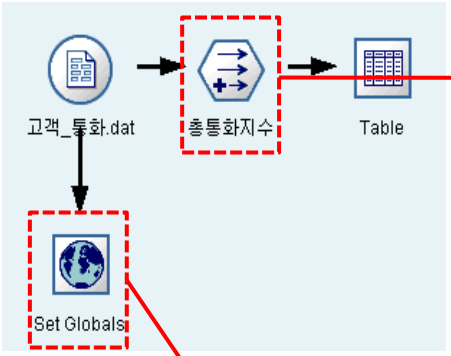


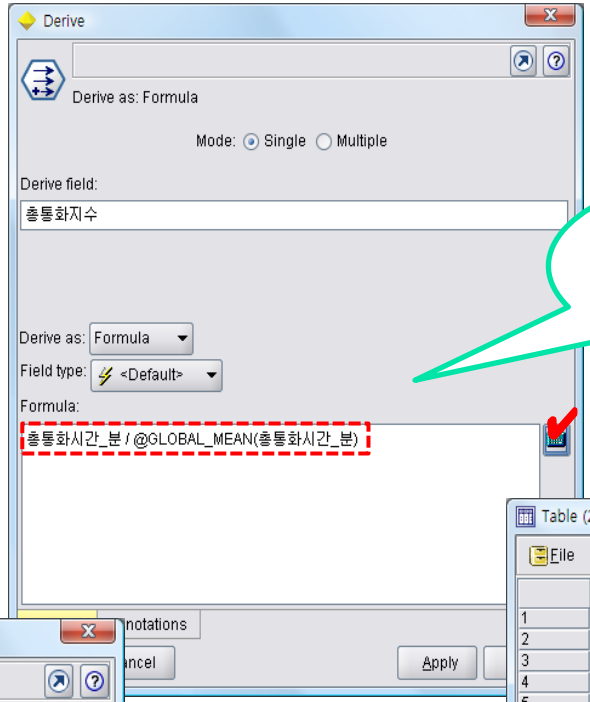


월평균수입 변수를 4분위수와 10분위수 두가지로 구간화함

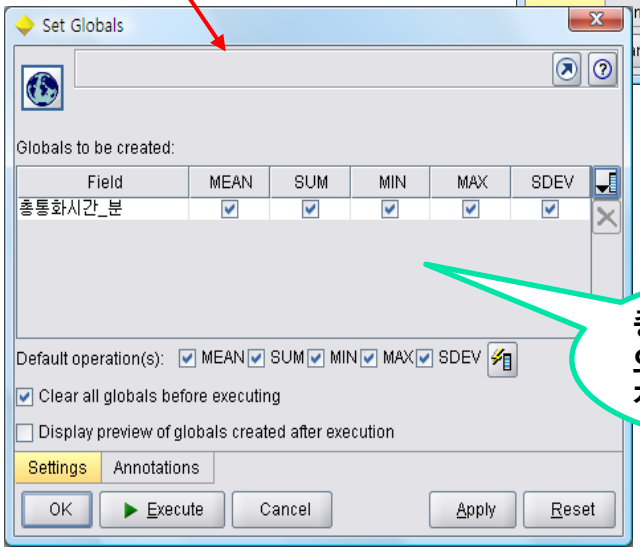
전역 값 설정

실습 DATA
- 고객_통화.dat

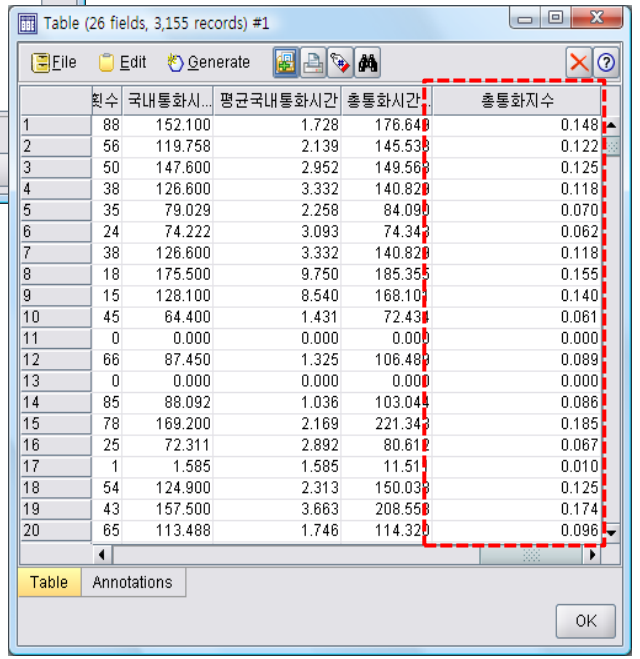




설정한 전역값을 파생 노드를 통해 새로운 변수 생성에 사용함



총통화시간_분의 전역값을 설정함

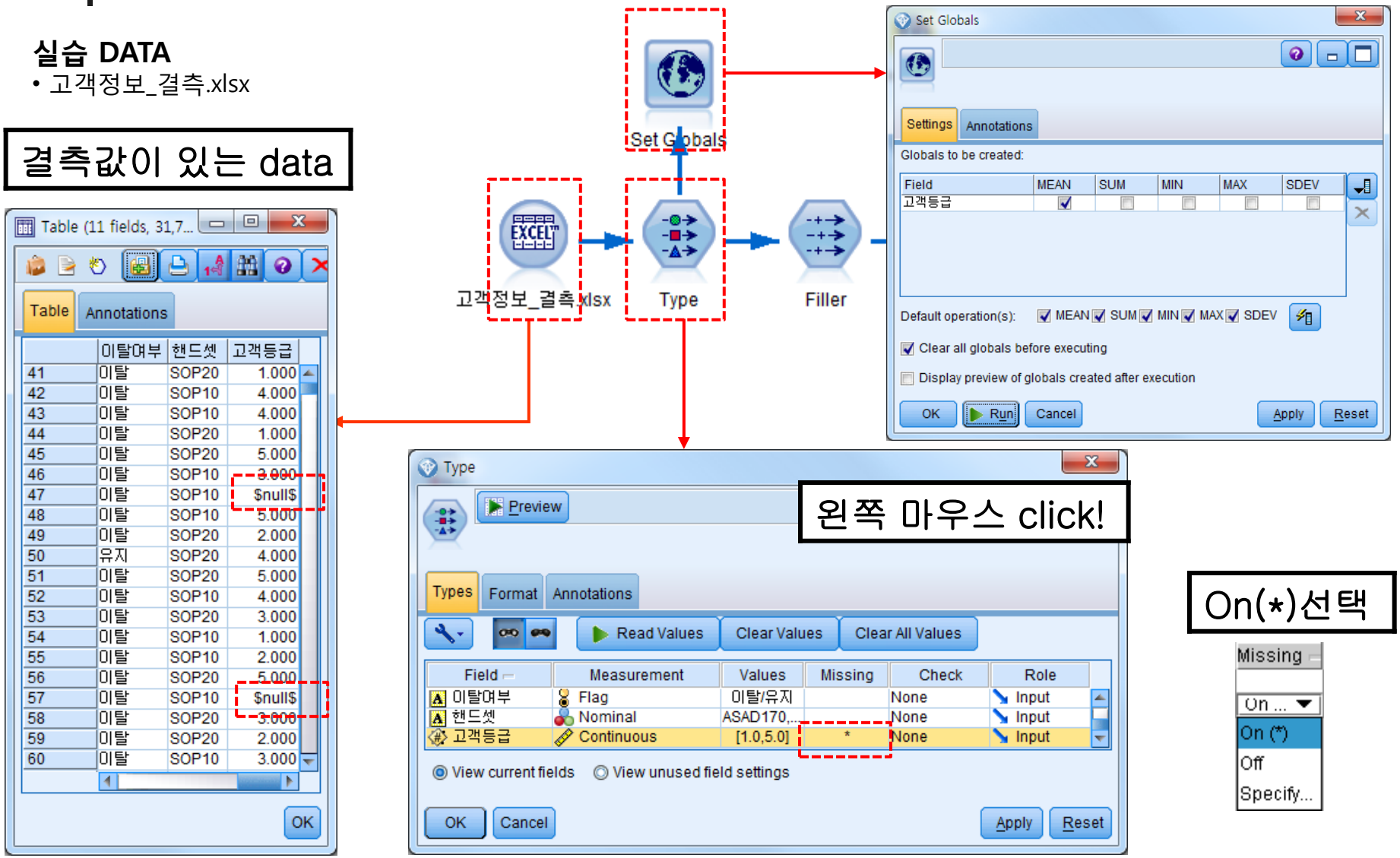


▪ **전역값(Filler) 노드**
: 필드의 통합 계산된 값을 memory에 저장하여 스트림내에서 자유롭게 사용할 수 있게 한다.

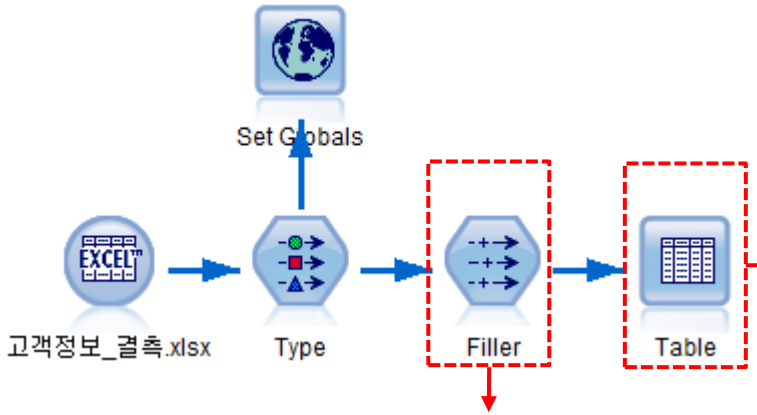
결측 값 처리 (1/2)

실습 DATA
- 고객정보_결측.xlsx

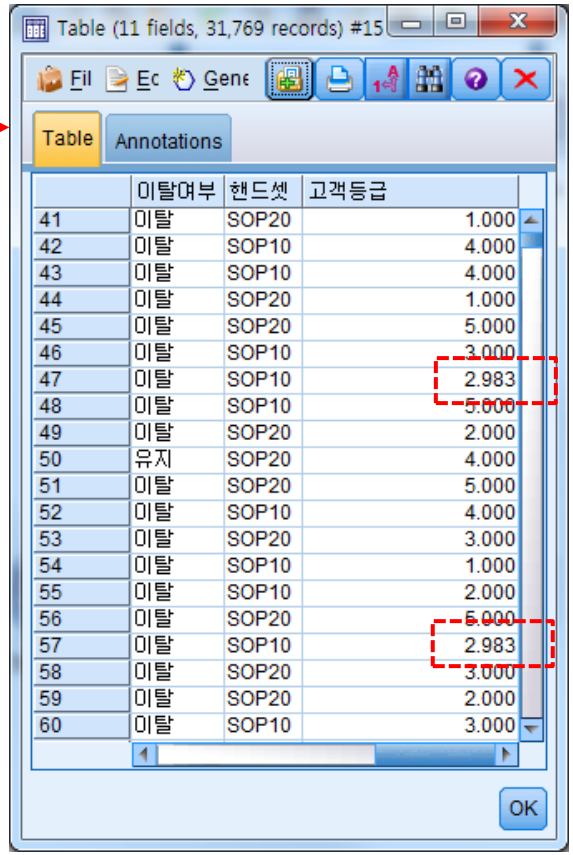
결측값이 있는 data



결측 값 처리 (2/2)

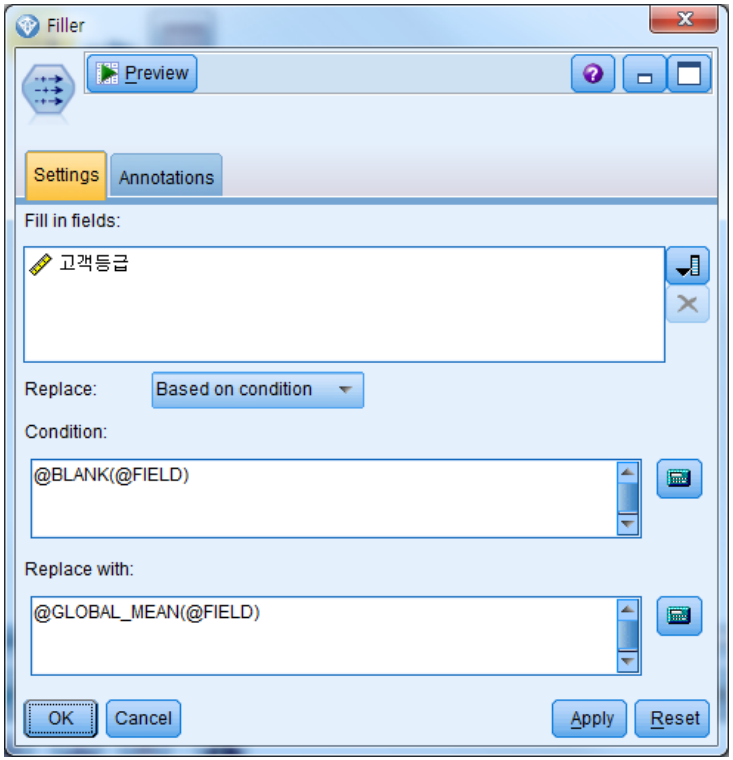


대체된 결측값

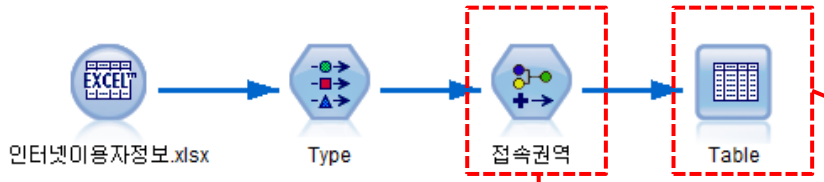


	이탈여부	핸드셋	고객등급
41	이탈	SOP20	1.000
42	이탈	SOP10	4.000
43	이탈	SOP10	4.000
44	이탈	SOP20	1.000
45	이탈	SOP20	5.000
46	이탈	SOP10	3.000
47	이탈	SOP10	2.983
48	이탈	SOP10	5.000
49	이탈	SOP20	2.000
50	유지	SOP20	4.000
51	이탈	SOP20	5.000
52	이탈	SOP10	4.000
53	이탈	SOP20	3.000
54	이탈	SOP10	1.000
55	이탈	SOP10	2.000
56	이탈	SOP20	5.000
57	이탈	SOP10	2.983
58	이탈	SOP20	3.000
59	이탈	SOP20	2.000
60	이탈	SOP10	3.000

'고객등급'의 평균값으로 결측값을 대체하는 대화상자

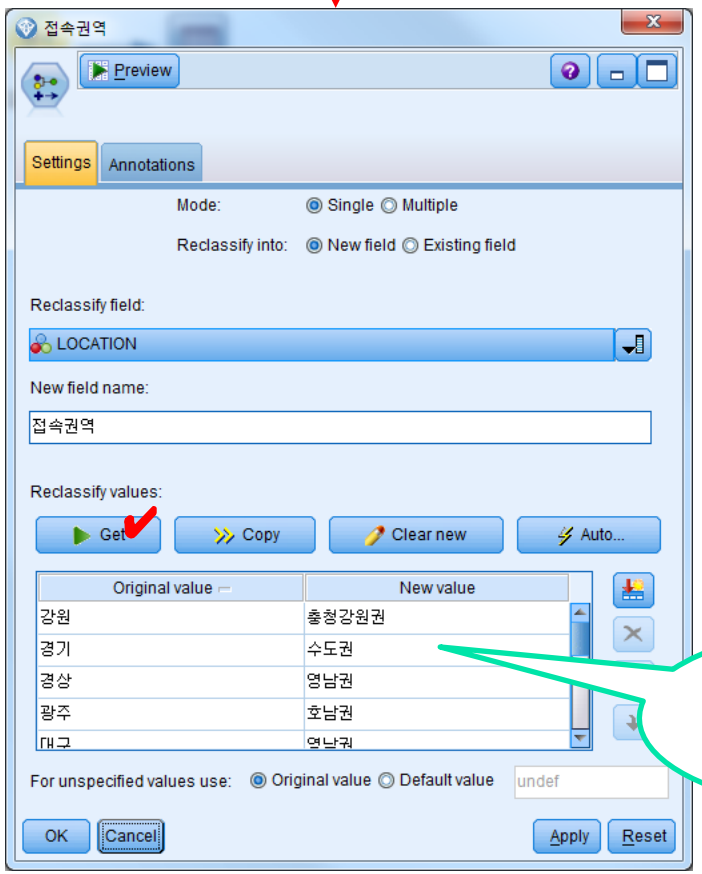


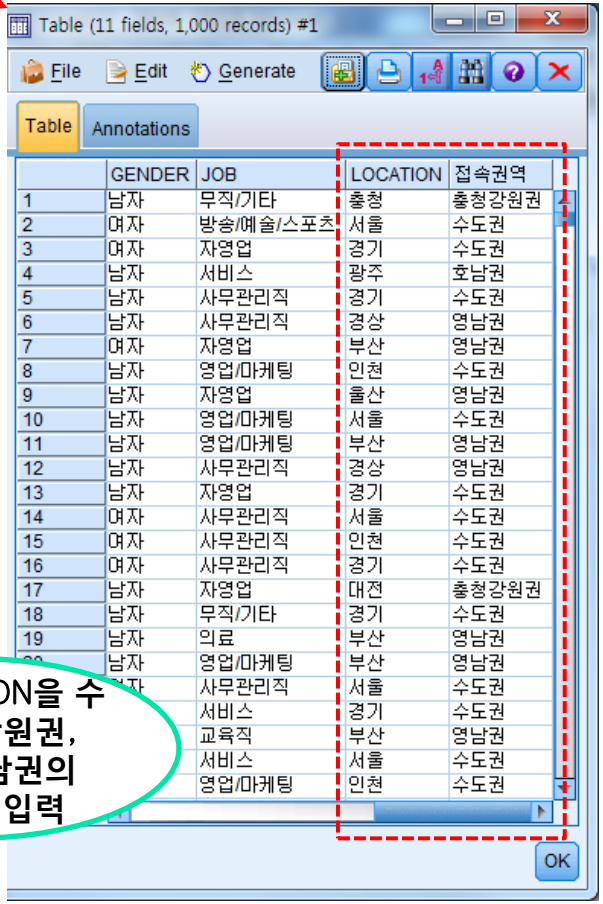
데이터 재분류



❖ **재분류(Reclassify) 노드**

이산형 변수에 대해 여러 가지 범주 값을 하나의 범주 값으로 바꾸거나 기존 값을 다른 값으로 바꿀 때 사용



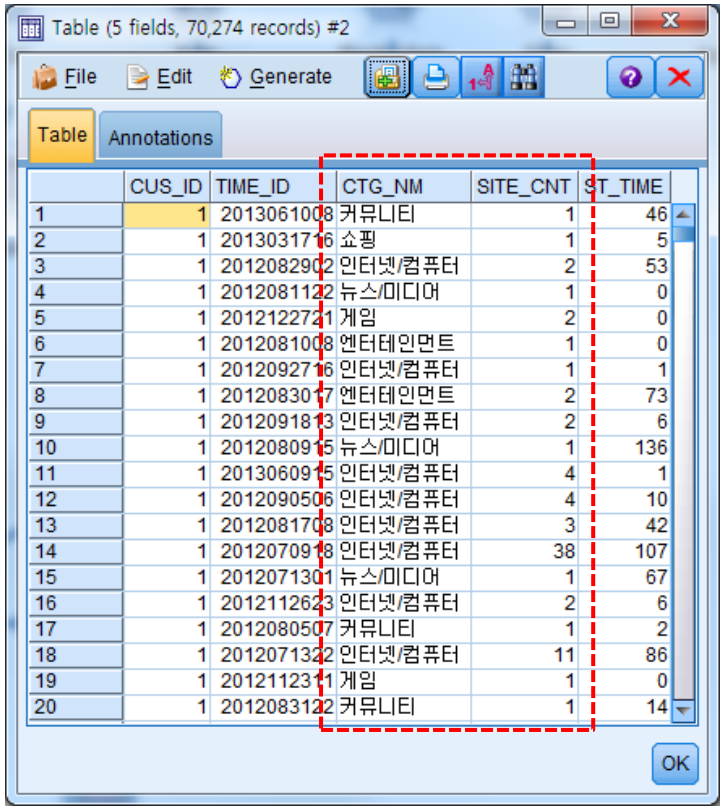


각 LOCATION을 수도권,충청강원권, 영남권, 호남권의 대권역으로 입력

데이터 재구성 (1/3)

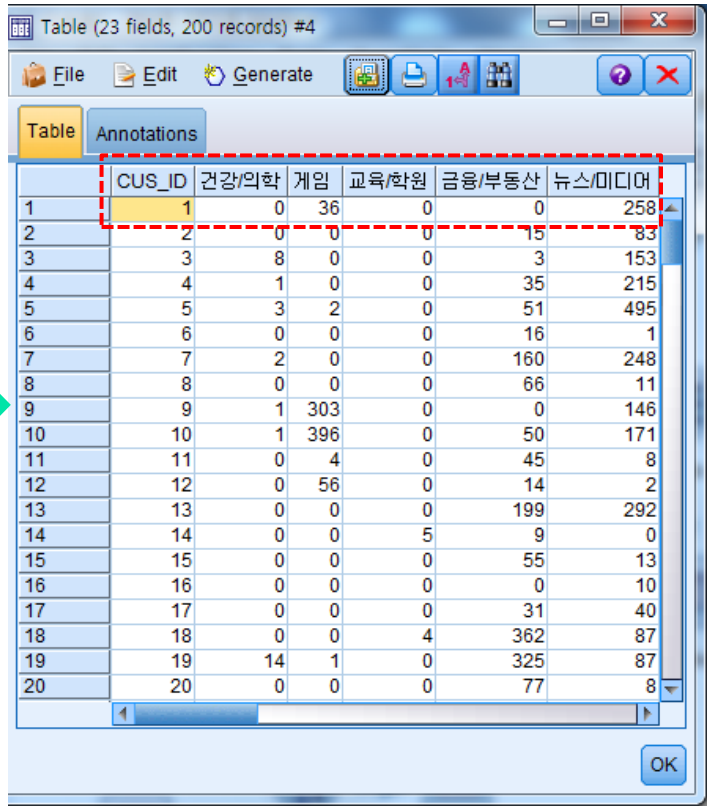
BEFORE

AFTER



	CUS_ID	TIME_ID	CTG_NM	SITE_CNT	ST_TIME
1	1	2013061008	커뮤니티	1	46
2	1	2013031716	쇼핑	1	5
3	1	2012082902	인터넷/컴퓨터	2	53
4	1	2012081122	뉴스/미디어	1	0
5	1	2012122721	게임	2	0
6	1	2012081008	엔터테인먼트	1	0
7	1	2012092716	인터넷/컴퓨터	1	1
8	1	2012083017	엔터테인먼트	2	73
9	1	2012091813	인터넷/컴퓨터	2	6
10	1	2012080915	뉴스/미디어	1	136
11	1	2013060915	인터넷/컴퓨터	4	1
12	1	2012090506	인터넷/컴퓨터	4	10
13	1	2012081708	인터넷/컴퓨터	3	42
14	1	2012070918	인터넷/컴퓨터	38	107
15	1	2012071301	뉴스/미디어	1	67
16	1	2012112623	인터넷/컴퓨터	2	6
17	1	2012080507	커뮤니티	1	2
18	1	2012071322	인터넷/컴퓨터	11	86
19	1	2012112311	게임	1	0
20	1	2012083122	커뮤니티	1	14

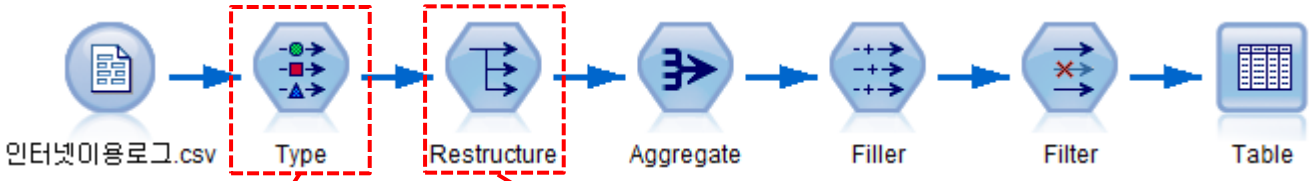
Restructuring

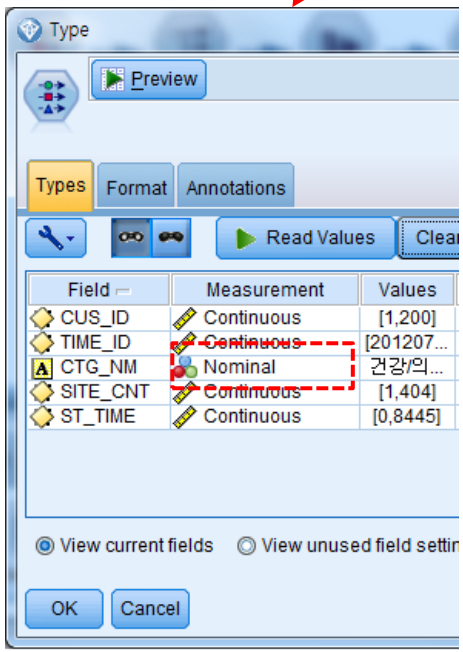


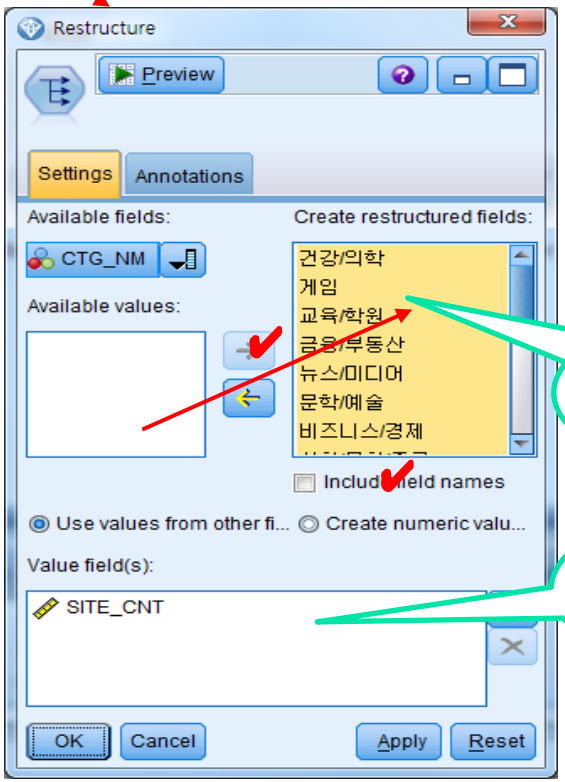
	CUS_ID	건강/의학	게임	교육/학원	금융/부동산	뉴스/미디어
1	1	0	36	0	0	258
2	2	0	0	0	15	83
3	3	8	0	0	3	153
4	4	1	0	0	35	215
5	5	3	2	0	51	495
6	6	0	0	0	16	1
7	7	2	0	0	160	248
8	8	0	0	0	66	11
9	9	1	303	0	0	146
10	10	1	396	0	50	171
11	11	0	4	0	45	8
12	12	0	56	0	14	2
13	13	0	0	0	199	292
14	14	0	0	5	9	0
15	15	0	0	0	55	13
16	16	0	0	0	0	10
17	17	0	0	0	31	40
18	18	0	0	4	362	87
19	19	14	1	0	325	87
20	20	0	0	0	77	8

데이터 재구성 (2/3)

❖ **재구성(Restructure) 노드**
필드 유형이 nominal 또는 flag인 필드의 값을 이용하여 다중 필드를 생성(SetToFlag 노드의 일반형).



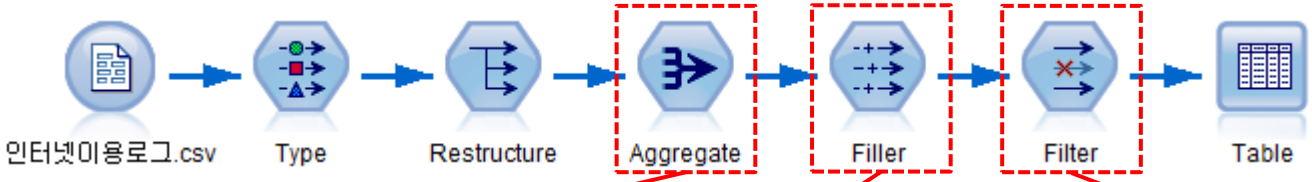


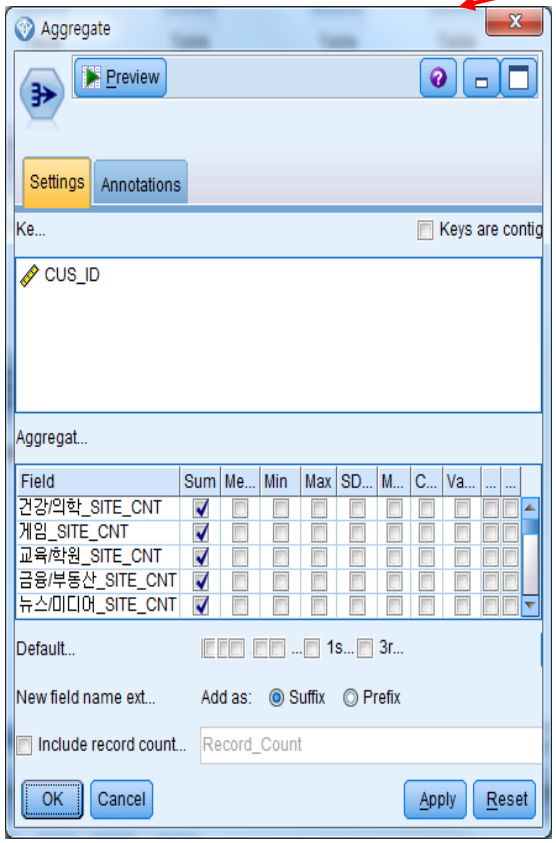


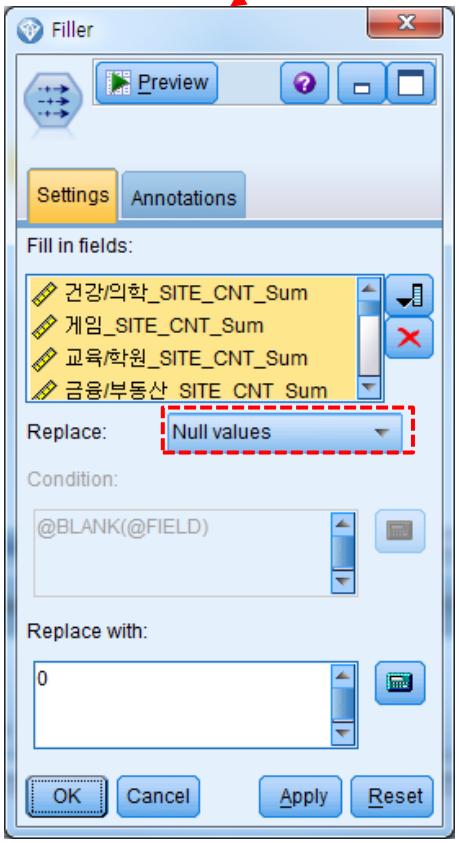
CTG_NM 필드의 값 -> 새로 생성할 필드명

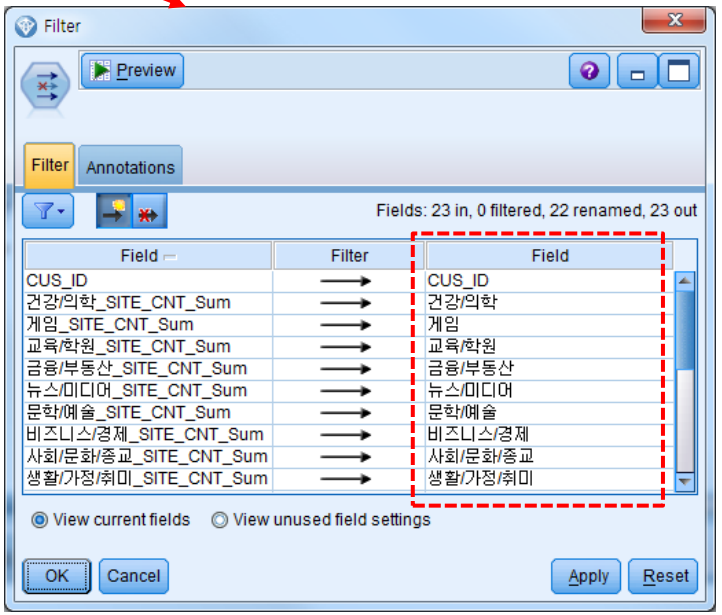
SITE_CNT 필드의 값으로 새로 생성되는 필드의 값을 채움.

데이터 재구성 (3/3)

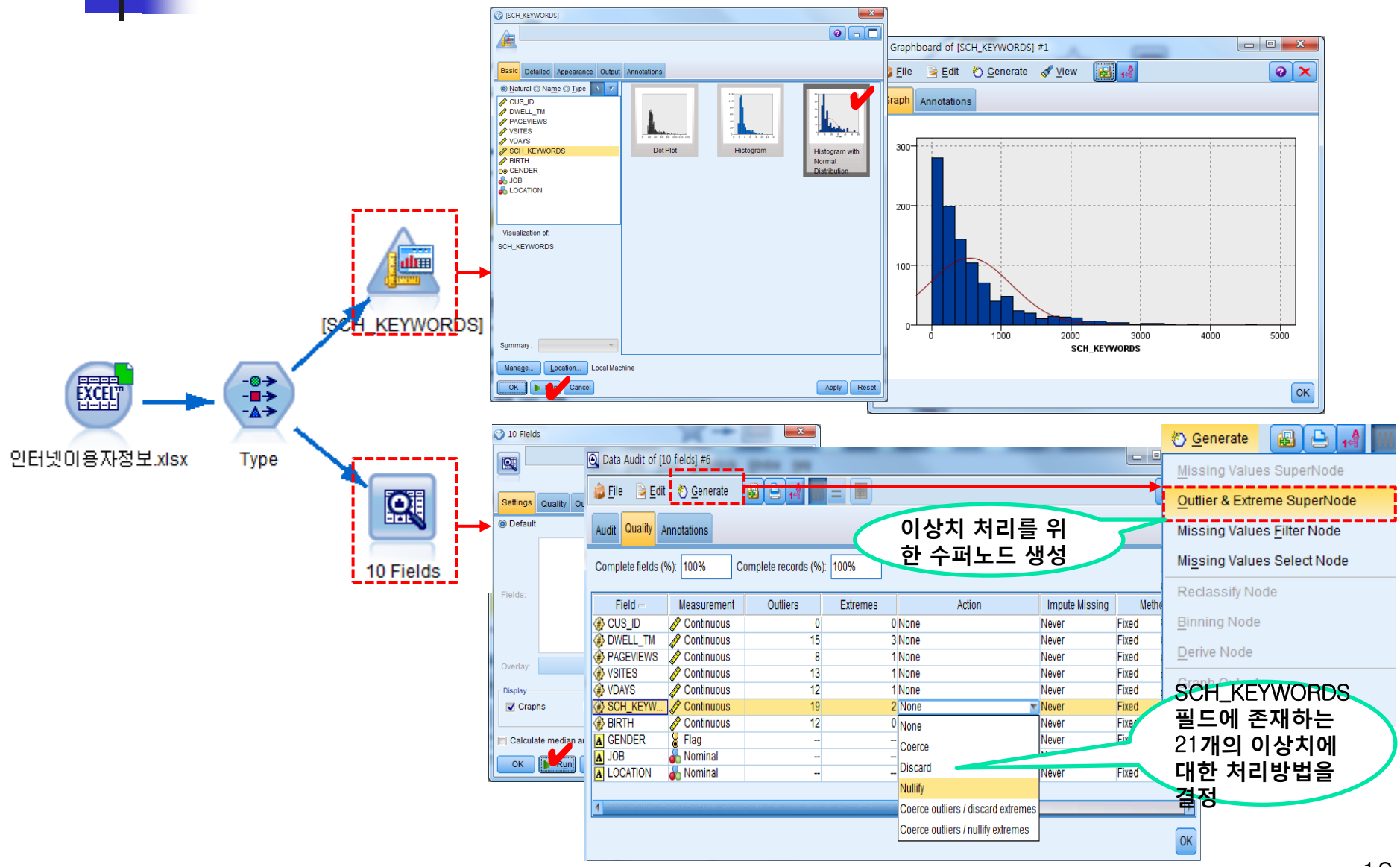

인터넷이용로그.csv
Type
Restructure
Aggregate
Filler
Filter
Table


Aggregate
Preview
Settings
Annotations
Ke...
Keys are contig
CUS_ID
Aggregat...
Field
Sum
Me...
Min
Max
SD...
M...
C...
Va...
건강/의학_SITE_CNT
게임_SITE_CNT
교육/학원_SITE_CNT
금융/부동산_SITE_CNT
뉴스/미디어_SITE_CNT
Default...
1s...
3r...
New field name ext...
Add as:
Suffix
Prefix
Include record count...
Record_Count
OK
Cancel
Apply
Reset


Filler
Preview
Settings
Annotations
Fill in fields:
건강/의학_SITE_CNT_Sum
게임_SITE_CNT_Sum
교육/학원_SITE_CNT_Sum
금융/부동산_SITE_CNT_Sum
Replace:
Null values
Condition:
@BLANK(@FIELD)
Replace with:
0
OK
Cancel
Apply
Reset


Filter
Preview
Filter
Annotations
Fields: 23 in, 0 filtered, 22 renamed, 23 out
Field
Filter
Field
CUS_ID
건강/의학_SITE_CNT_Sum
게임_SITE_CNT_Sum
교육/학원_SITE_CNT_Sum
금융/부동산_SITE_CNT_Sum
뉴스/미디어_SITE_CNT_Sum
문학/예술_SITE_CNT_Sum
비즈니스/경제_SITE_CNT_Sum
사회/문화/종교_SITE_CNT_Sum
생활/가정/취미_SITE_CNT_Sum
View current fields
View unused field settings
OK
Cancel
Apply
Reset

이상치(Outlier) 처리 (1/2)



이상치(Outlier) 처리 (2/2)

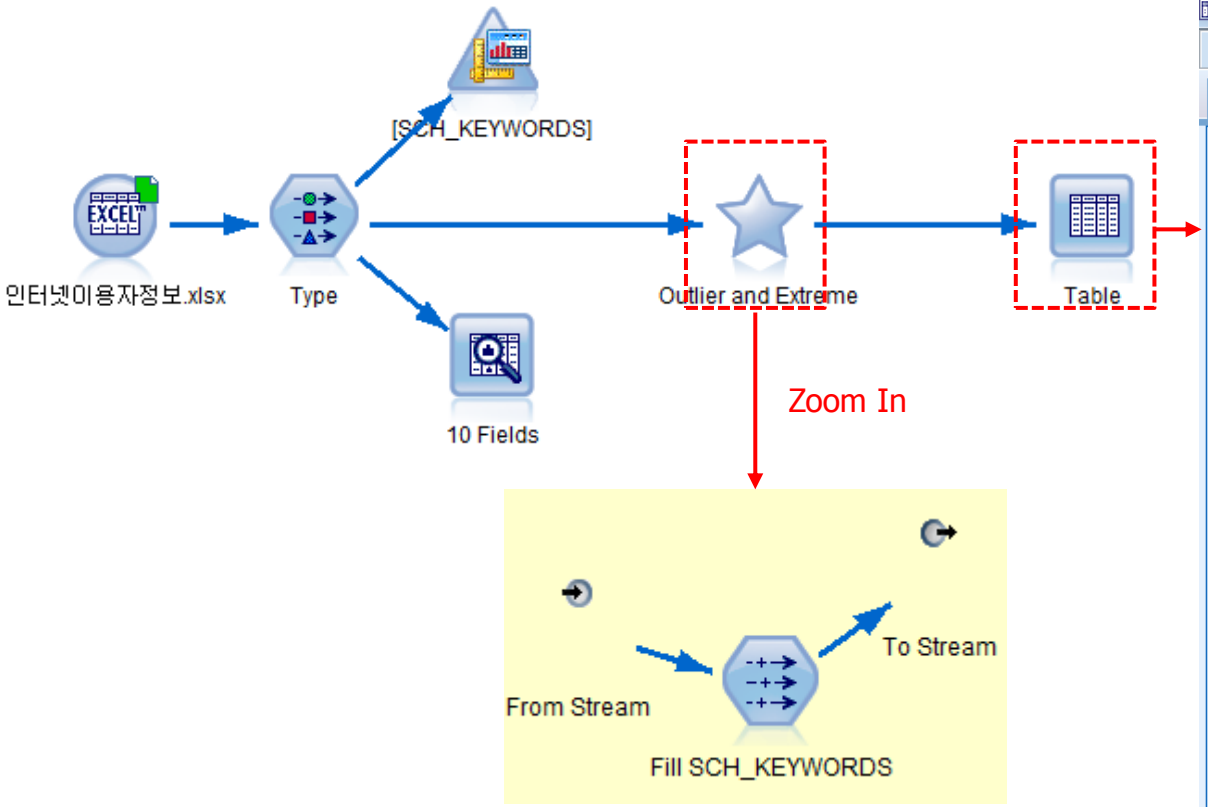


Table (10 fields, 1,000 records) #2

File Edit Generate

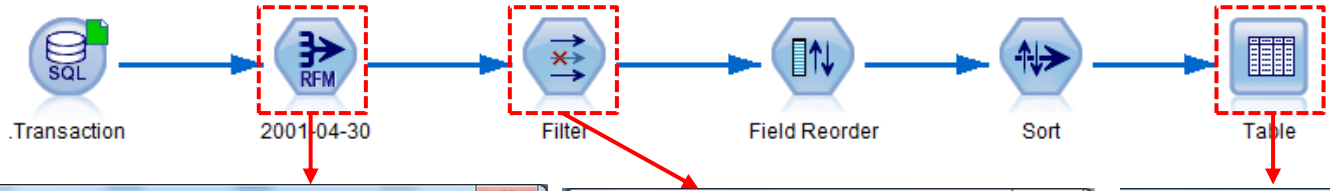
Table Annotations

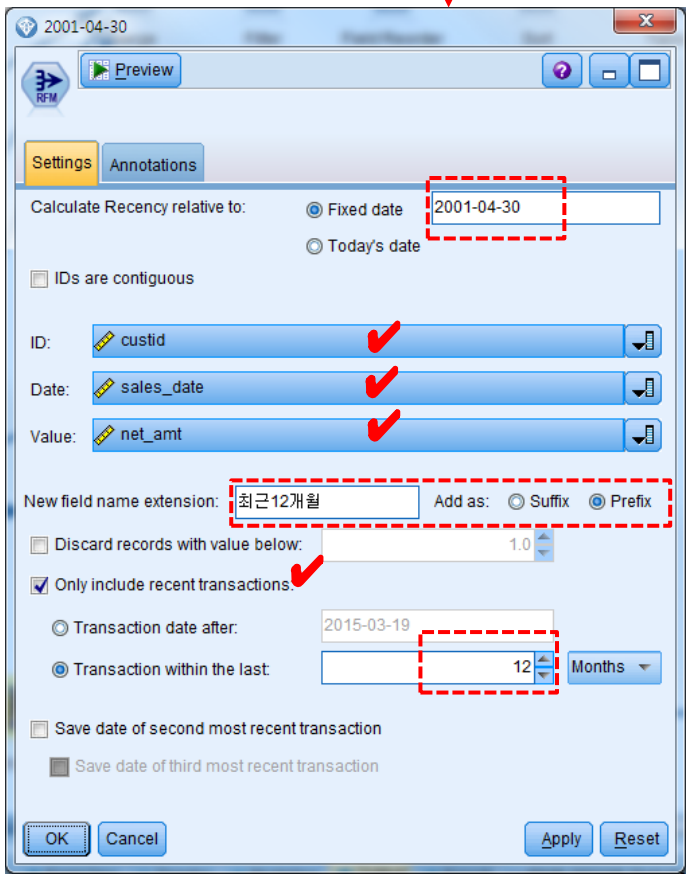
	CUS_ID	DWELL_...	PAGEVIEWS	VSITES	VDAYS	SCH_KEYWORDS
1	1.000	357978.0...	13656.000	201.000	205.0...	399.000
2	2.000	836048.0...	25498.000	573.000	105.0...	1346.000
3	3.000	388378.0...	10413.000	270.000	105.0...	329.000
4	4.000	275572.0...	12831.000	166.000	107.0...	134.000
5	5.000	501922.0...	13267.000	328.000	151.0...	511.000
6	6.000	56543.000	1972.000	85.000	79.000	62.000
7	7.000	573113.0...	20063.000	228.000	243.0...	87.000
8	8.000	101640.0...	5871.000	316.000	173.0...	108.000
9	9.000	535910.0...	22054.000	438.000	255.0...	464.000
10	10.000	1222047....	43544.000	236.000	116.0...	290.000
11	11.000	90562.000	5582.000	450.000	123.0...	82.000
12	12.000	68144.000	4376.000	68.000	113.0...	12.000
13	13.000	426988.0...	10672.000	129.000	153.0...	84.000
14	14.000	644274.0...	21572.000	218.000	100.0...	658.000
15	15.000	595743.0...	25967.000	330.000	174.0...	656.000
16	16.000	90895.000	3532.000	159.000	100.0...	469.000
17	17.000	172435.0...	7284.000	175.000	101.0...	45.000
18	18.000	518476.0...	11597.000	250.000	201.0...	193.000
19	19.000	329314.0...	12087.000	260.000	101.0...	242.000
20	20.000	125222.0...	7163.000	247.000	123.0...	195.000
21	21.000	102183.0...	3400.000	83.000	119.0...	29.000
22	22.000	302502.0...	12772.000	241.000	111.0...	309.000
23	23.000	489285.0...	23935.000	793.000	119.0...	956.000
24	24.000	421206.0...	12917.000	197.000	130.0...	212.000
25	25.000	385731.0...	36266.000	345.000	206.0...	$null$
26	26.000	530405.0...	16921.000	316.000	111.0...	$null$
27	27.000	546656.0...	21490.000	428.000	139.0...	960.000
28	28.000	592353.0...	23327.000	580.000	127.0...	1673.000
29	29.000	110400.0...	4760.000	201.000	100.0...	63.000
30	30.000	689042.0...	17056.000	286.000	112.0...	229.000
31	31.000	631398.0...	33939.000	377.000	201.0...	1825.000
32	32.000	1181614....	36883.000	380.000	108.0...	840.000
33	33.000	394670.0...	17793.000	302.000	107.0...	309.000
34	34.000	609035.0...	16023.000	369.000	159.0...	1100.000
35	35.000	310133.0...	8351.000	181.000	146.0...	68.000

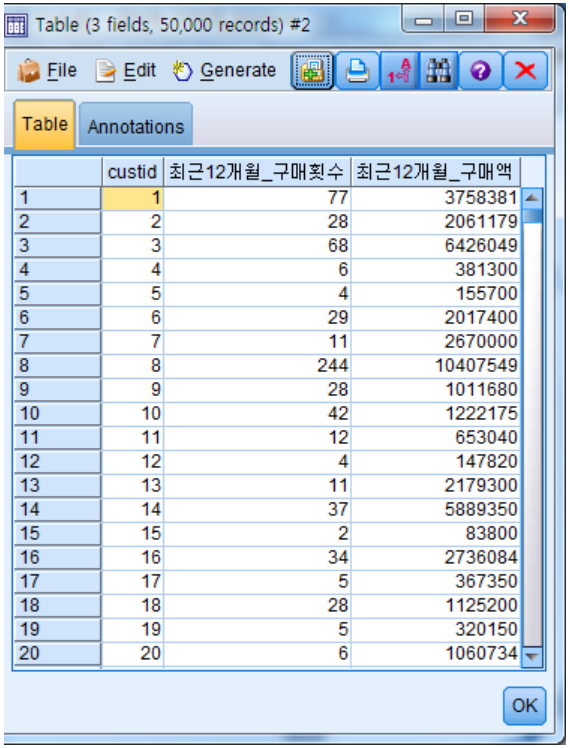
OK

파생변수 – 기간별 구매 금액•횟수•여부 (1/2)

- H백화점 고객의 최근 12개월 구매금액 및 구매횟수에 대한 변수 생성



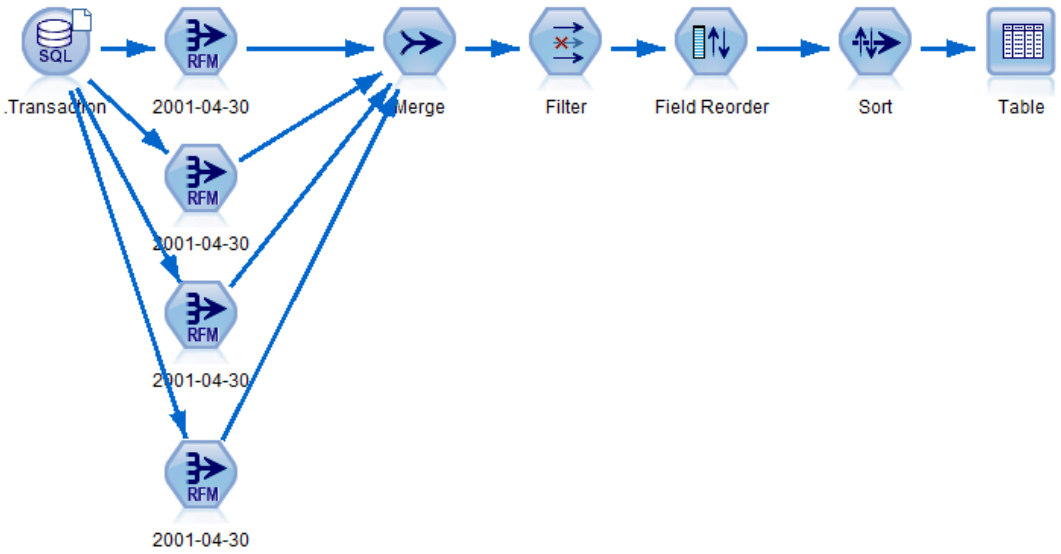




	custid	최근12개월_구매횟수	최근12개월_구매액
1	1	77	3758381
2	2	28	2061179
3	3	68	6426049
4	4	6	381300
5	5	4	155700
6	6	29	2017400
7	7	11	2670000
8	8	244	10407549
9	9	28	1011680
10	10	42	1222175
11	11	12	653040
12	12	4	147820
13	13	11	2179300
14	14	37	5889350
15	15	2	83800
16	16	34	2736084
17	17	5	367350
18	18	28	1125200
19	19	5	320150
20	20	6	1060734

파생변수 – 기간별 구매 금액•횟수•여부 (2/2)

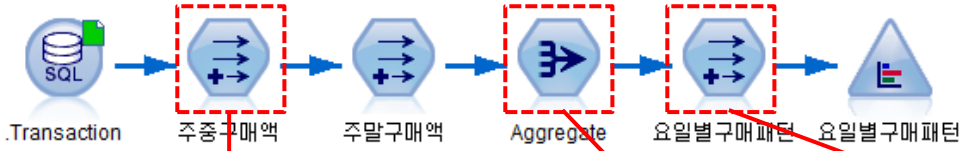
- H백화점 고객의 최근 1개월, 3개월, 6개월, 12개월 구매금액 및 구매 횟수에 대한 변수를 생성하시오.

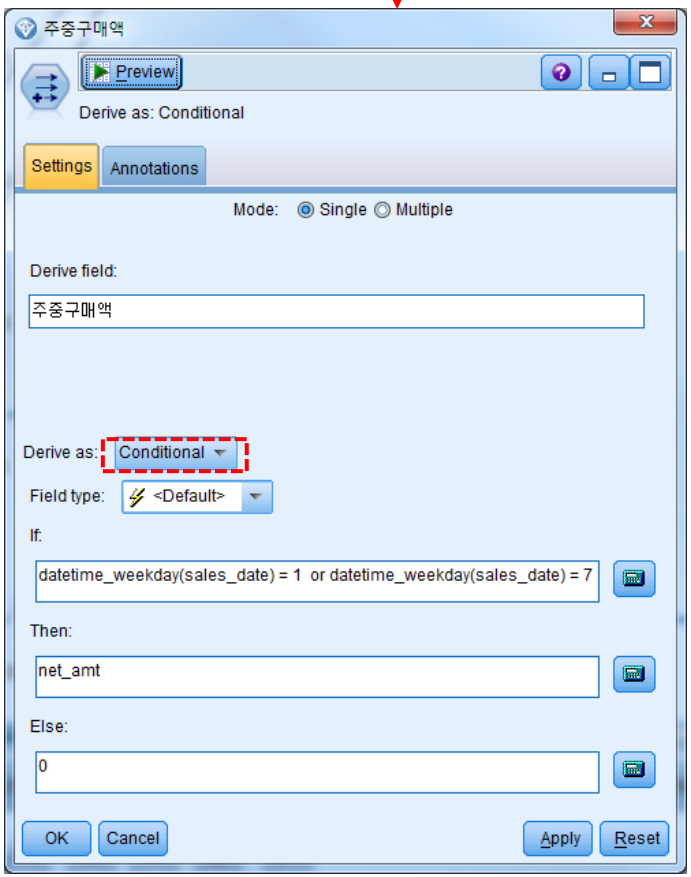


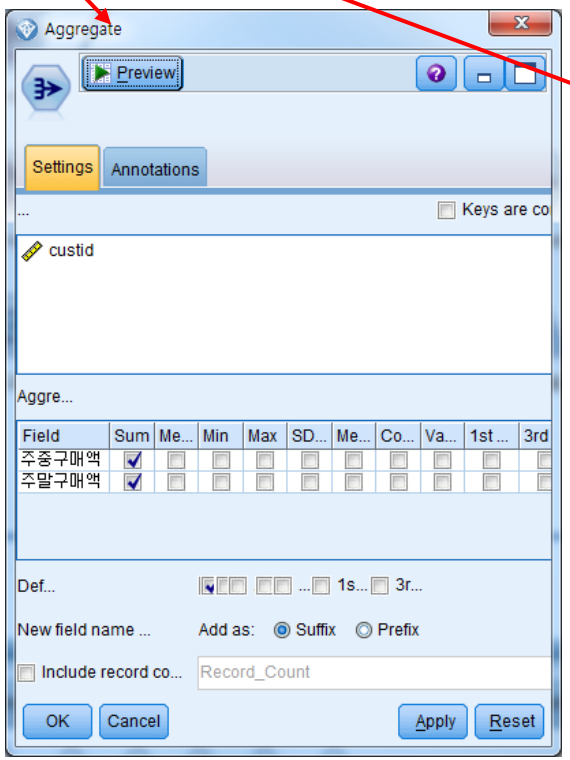
- 구매액 또는 구매횟수 기준으로 볼 때 매출이 줄어드는 고객은 누구인가? 반대로 매출이 증가하고 있는 고객은?

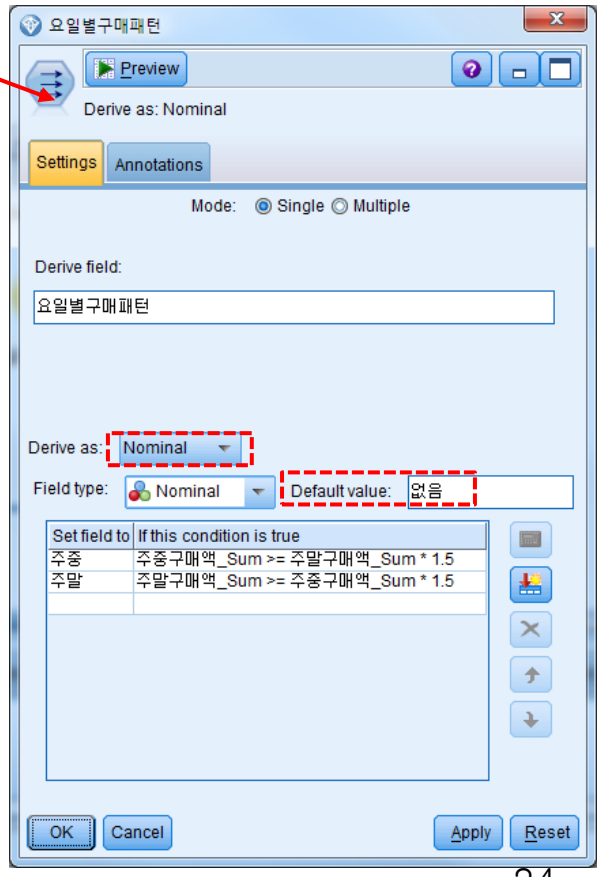
파생변수 – 요일별 구매패턴

- H백화점 고객의 주중•주말 구매패턴에 대한 변수 생성



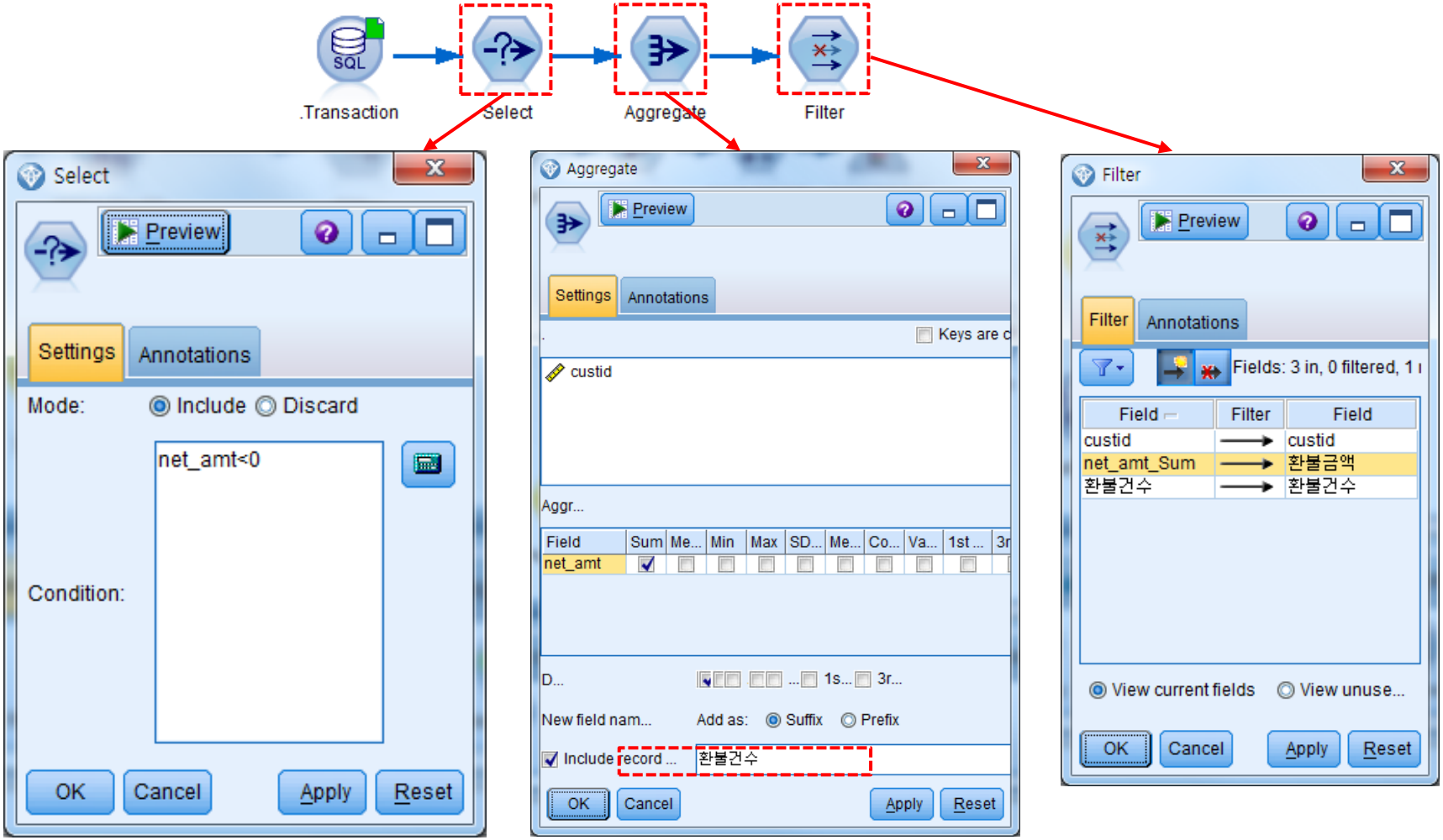






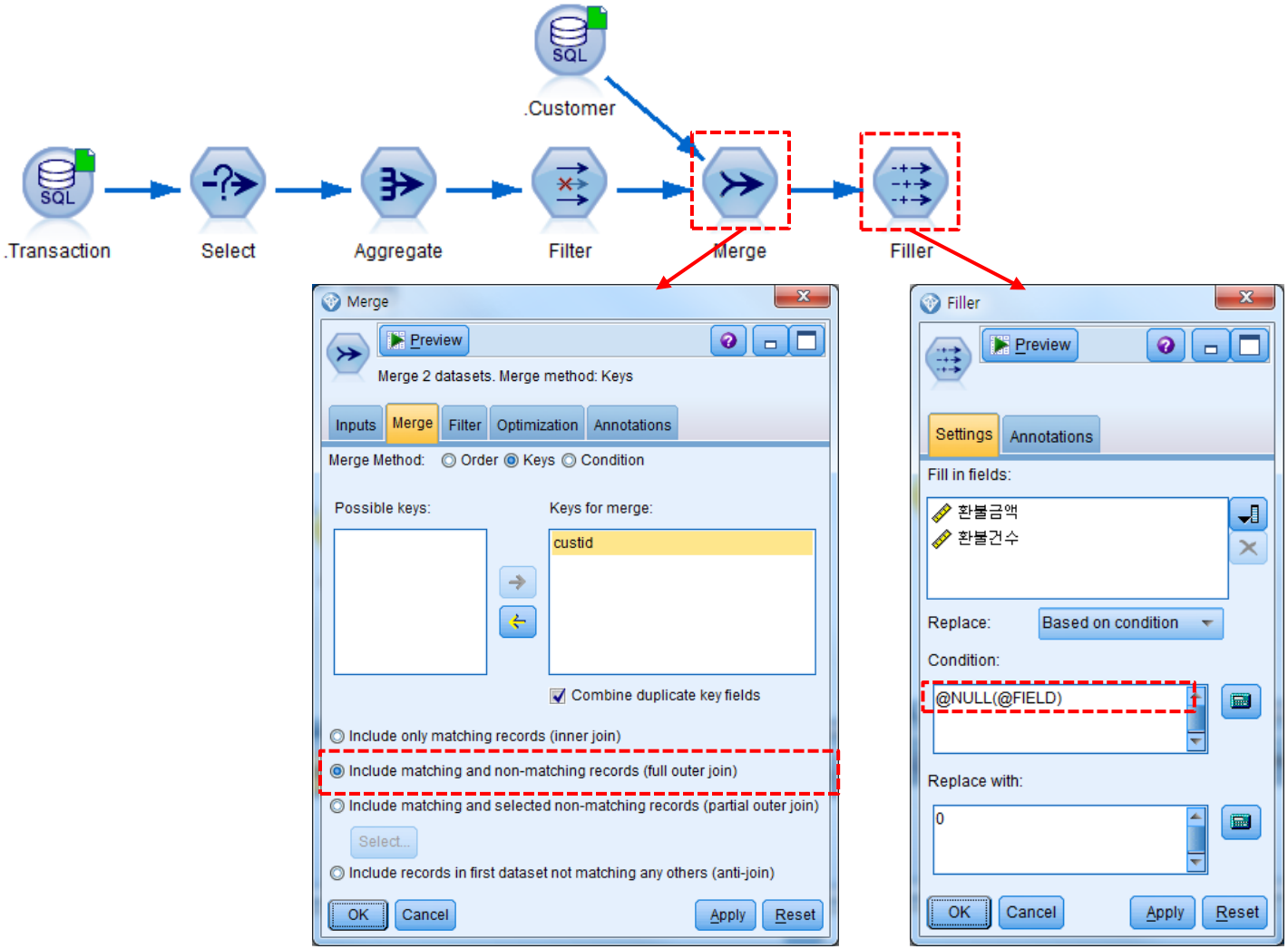
파생변수 – 환불행태 (1/2)

- H백화점 고객의 환불행태(금액, 건수)에 대한 변수 생성



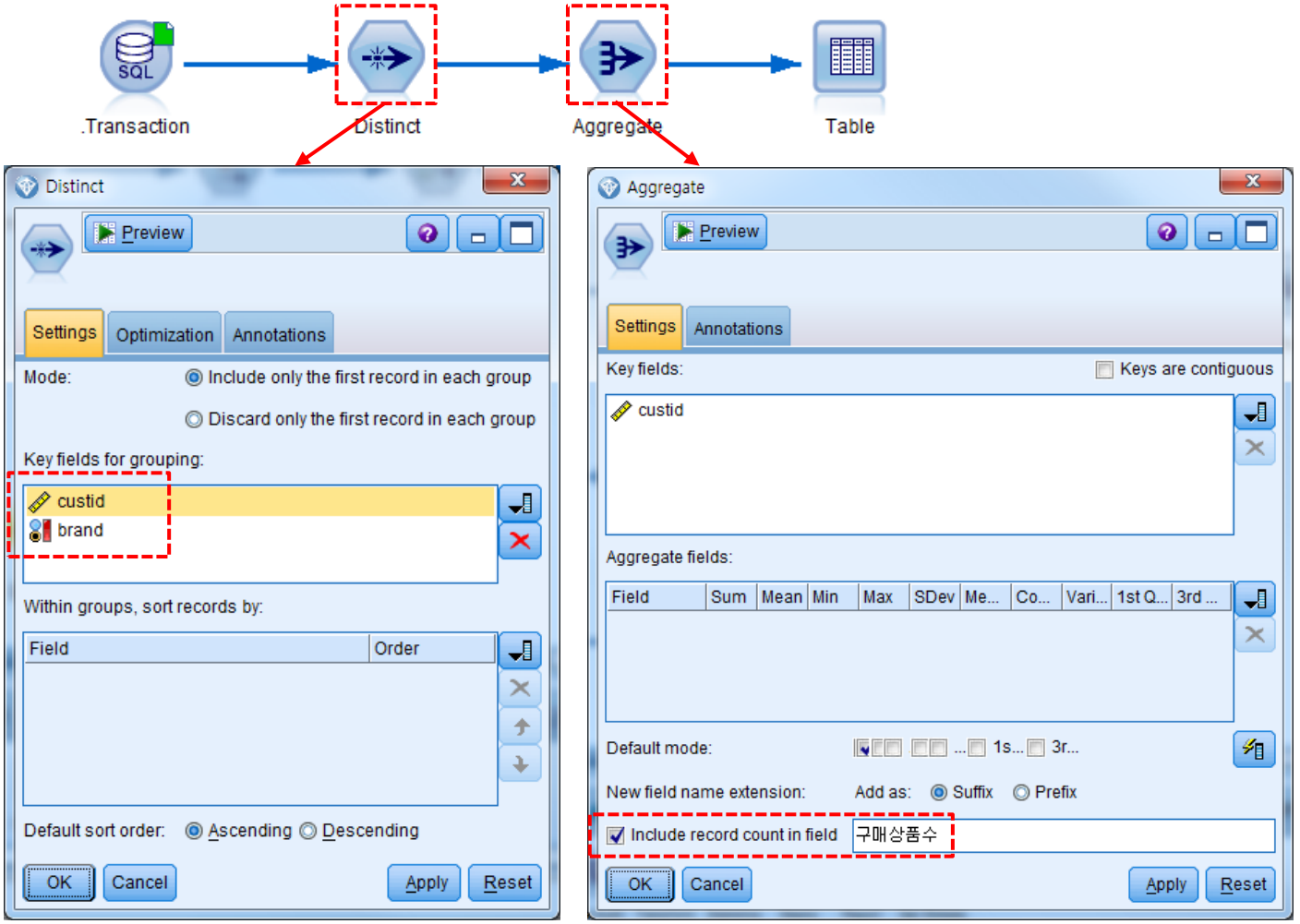
파생변수 – 환불행태 (2/2)

- H백화점 고객속성에 환불행태(금액, 건수) 변수를 결합



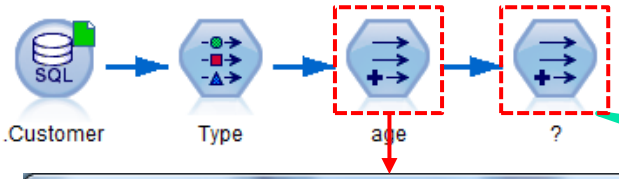
파생변수 – 구매상품 다양성

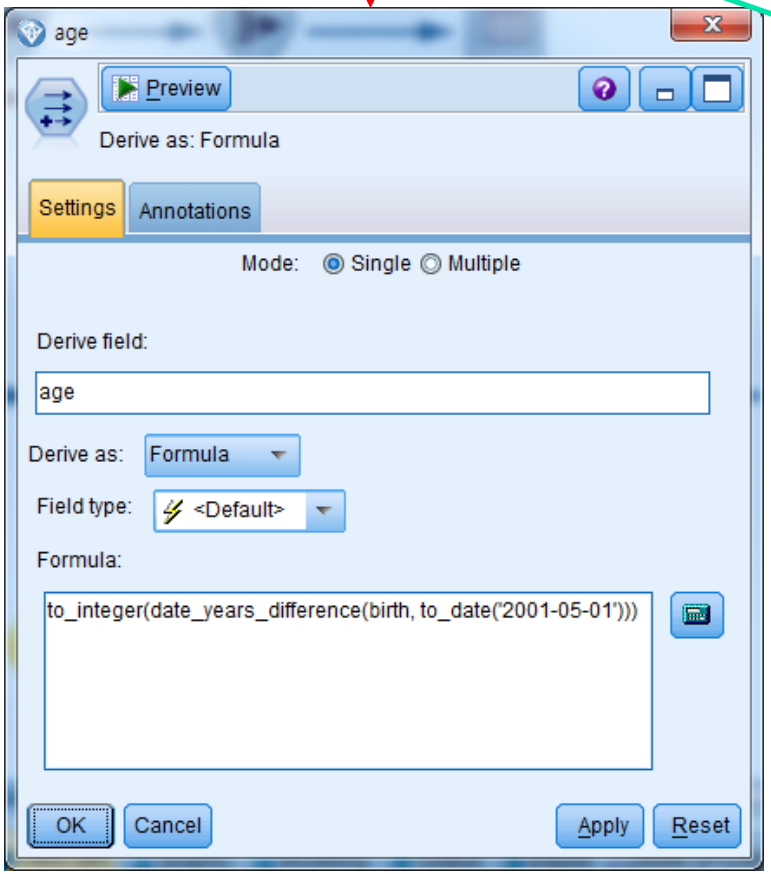
- H백화점 고객의 구매상품 다양성에 대한 변수 생성



파생변수 – 연령대

- H백화점 고객의 생일로부터 특정시점의 연령대를 계산하시오.



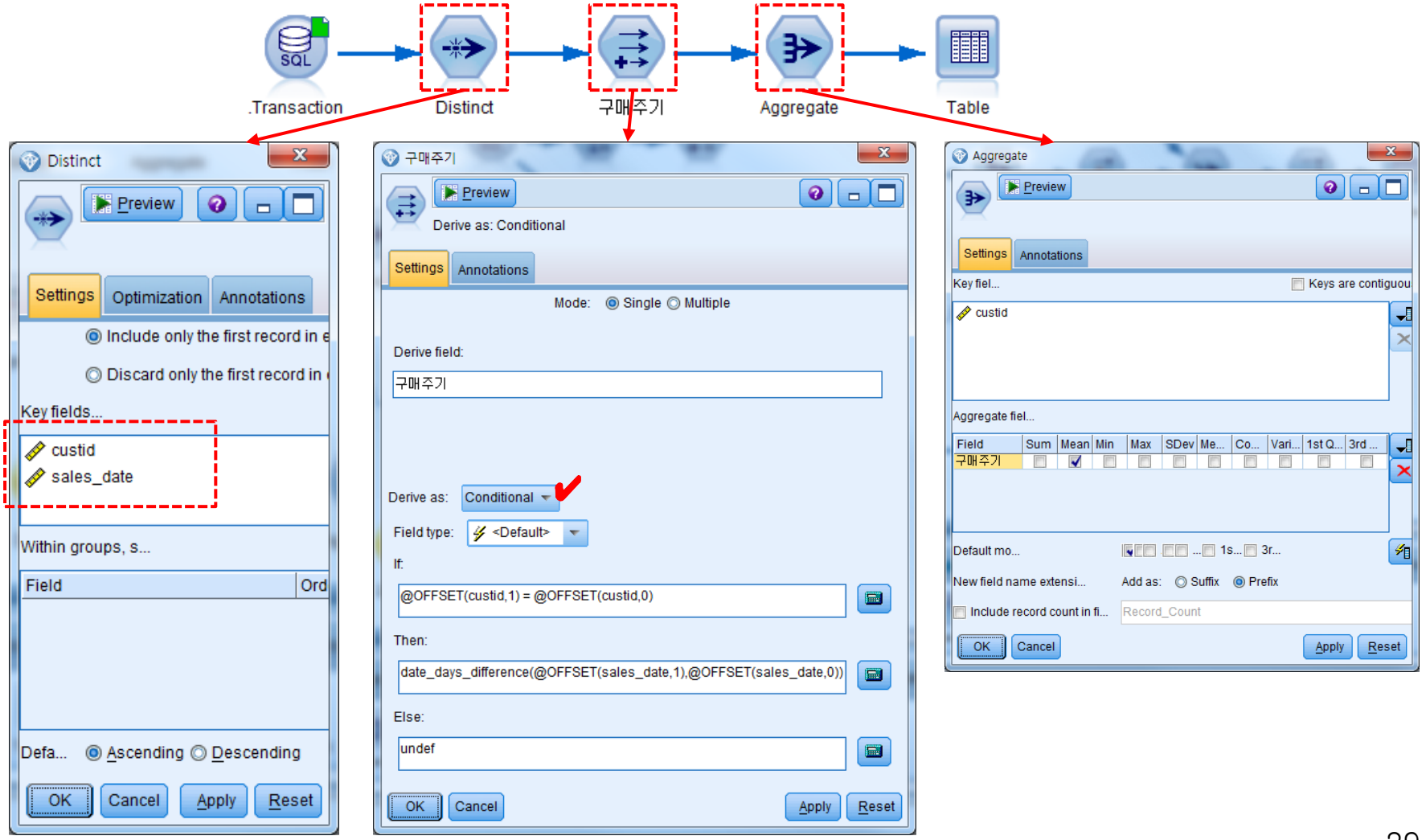


10대(이하),20대,30대,40대,50대,60대,70대(이상)으로 구간을 나눔

주의) 이상치 처리가 필요함.

파생변수 – 구매주기

- H백화점 고객의 평균구매주기(Average Purchasing Interval)를 계산



개인과제 #1 –6월5일까지 e메일제출

H백화점 데이터를 이용하여 아래와 같은 파생변수를 만드시오.

- 가격 선호도 변수
- 시즌 선호도 변수
- 상품별 구매 금액/횟수/여부 변수
- 상품별 구매순서 변수
- 주 구매상품 변수
- 휴면/이탈 가망 변수
  - Ex) If 평균구매주기 < 최종구매경과(현재 – 마지막 구매시점) Then “이탈”